Interval Privacy: A New Data Privacy Framework

Jie Ding  
School of Statistics  
University of Minnesota  
MN, 55414, USA  

Bangjun Ding  
School of Statistics and Finance  
East China Normal University  
Shanghai 200241, China

Abstract

The emerging public awareness and government regulations of data privacy motivate new paradigms of collecting and analyzing data transparent and acceptable to data owners. We present a novel concept of data privacy and corresponding data formats, mechanisms, and tradeoffs for privatizing data before they are collected and analyzed. The privacy, named Interval Privacy, enforces the raw data conditional distribution on the privatized data to be the same as its unconditional distribution over a nontrivial support set. The proposed privatizing mechanism is based on interval censoring techniques, where a set of points is recorded as a set of intervals containing them. To our best knowledge, this is the first work that advocates the use of random intervals to mask raw data and systematically enhance data privacy. A unique feature of interval mechanisms is that they obfuscate the truth but not distort it. Various theoretical aspects of the proposed privacy are studied. Moreover, in the context of supervised learning, we propose a novel method such that existing supervised learning algorithms designed for point-valued data could be directly applied to learn from interval-valued data.

Keywords: data collection, interval privacy, interval mechanism, local privacy, regression.

1. Introduction

With new and far-reaching laws such as the General Data Protection Regulation (Voigt and Von dem Bussche, 2017), and frequent headlines of large-scale data breaches, there has been a growing societal concern about how personal data is collected and used (Evans et al., 2015; Cross and Cavallaro, 2020). Consequently, data privacy has been an increasingly important factor in designing signal processing and machine learning services (Google, 2019; Facebook, 2020). In this paper, we will address the following scenario often seen in practice. Suppose that Alice is the agent who creates and holds raw data, which will be collected by another agent Bob. On the one hand, Alice may not trust Bob or the transmission channel to Bob. On the other hand, Bob is interested in population-wide inference using the data provided by Alice and many other individuals, but not necessarily the exact value held by Alice.

The above learning scenario is quite common in, e.g., Machine-Learning-as-a-Service cloud services (Alabbadi, 2011; Ribeiro et al., 2015), multi-organizational Assisted Learning (Xian et al., 2020), survey-based inferences (Couper et al., 2001), and information fusion (Wang et al., 2018; Sun and Tay, 2019b). The tradeoffs between individual-level data privacy and population-level learning utility have motivated active research on what
Local data privacy is generally referred to as local data privacy across fields such as data mining (Evfimievski et al., 2003), security (Kasiviswanathan et al., 2011), statistics (Duchi et al., 2018), and information theory (du Pin Calmon and Fawaz, 2012; Sun et al., 2016). Local data privacy’s general goal is to suitably process raw data (often by randomizations) before its collection and evaluate privacy-utility tradeoffs through an appropriate framework. For example, a popular framework is the local differential privacy (Evfimievski et al., 2003; Kasiviswanathan et al., 2011), where Alice’ data $Y$ is suitably randomized (often by adding Laplacian noises) into $Z$ so that the ratio of conditional densities $p(z \mid y_1)/p(z \mid y_2)$ is bounded within $[e^{-\alpha}, e^\alpha]$ regardless of $z, y_1, y_2$. The value of $\alpha$ controls the level of privacy and utility. More discussions on related work will be included later in this paper.

In this work, we propose a novel concept of local privacy named interval privacy for protecting data collected for further inferences. The main idea is to enforce privacy in such a way that the distribution of the raw data conditional on its privatized data remains the same (up to a normalizing constant) on a large support set. In other words, no additional information is gained except that the support of the data becomes narrow. Accompanying the notion of interval privacy, we use the size (in a measure-theoretic sense) of the conditional support to quantify the level of privacy. The size, named as privacy coverage, enables a natural interpretation of the amount of ambiguity exposed to a potential adversary. We will discuss that interval privacy is neither a generalization nor a specialization of local differential privacy.

We then introduce a general privacy mechanism for realizing such privacy in practice. The main idea is to generate random intervals that partition the observation space for each raw data, and report which interval the data falls. We will also introduce a measure of privacy level named “privacy coverage”, based on the measure-theoretic interval size. This definition of privacy naturally motivates a new form of disclosing and collecting sensitive data, representing it as an interval instead of using a point. For example, a sensor’s accurate distance with the target $y = 10$ (in meters) is privatized by first generating a random threshold, say 20, and then publicizing the corresponding interval $(-\infty, 20]$; or an individual’s $60k$ salary (in dollars) is privatized by first generating random thresholds, say $41k$ and $85k$, and then reporting the interval $(41k, 85k]$. The random thresholds can be generated from any distribution known to the data collector, e.g., Gaussian, Logistic, and Uniform distributions. It is worth noting that an interval is not necessarily symmetric around the underlying raw data value. More generally, we will advocate the independence between thresholds and raw data so that adversaries cannot accurately decode the raw data.

Figure 1 illustrates $Y$ and its private counterpart in a dataset that will be used in the experimental section. From the figure, each individual’s privacy coverage describes the ‘interval width’ or level of ambiguity. For example, the data with 97% coverage is less private than the 99% one, which is in line with the perception that $Y < 82.2$ tells more information than $Y < 85.4$ does. We will show several fundamental properties of the proposed privacy mechanism to render its broad applicability. These include the composition property that characterizes the level of overall privacy degradation in the presence of multiple queries to the same data, robustness to pre-processing, and robustness to post-processing.

Though this is the first statistical framework that advocates random intervals for local data privacy, interval-valued data has been extensively studied in survival analysis (Sun, 2007). The theoretical development in survival analysis may be readily applied to data-
Figure 1: A snapshot of the database where the life expectancy (in years), $Y$, is privatized into a random interval, and the privacy coverage for each data item is provided. The average privacy coverage for the whole database is 60.3%.

private learning. For example, under mild conditions, it has been shown that the distribution function can be consistently estimated from interval observations using nonparametric maximum likelihood estimation (Groeneboom and Wellner, 1992), or a mixture of interval and point observations using Kaplan-Meier estimation (Kaplan and Meier, 1958; Turnbull, 1976). We will demonstrate the use of interval-private data for several learning tasks, including moment estimation, functional estimation, and supervised regression. For the particular case of supervised regression where the response labels are privatized, estimating the data-generating function is highly nontrivial when the number of predictors is large. We propose an algorithm to estimate regression functions using the interval privatized data. The main idea is to construct a surrogate response variable with the same expectation as the original response conditional on the predictors.

More specifically, the contributions of this paper are as follows.

- We propose a new perspective of data privacy named interval privacy, where raw data information is collected or released only though its feasible range. We show that interval privacy can be naturally implemented through interval mechanisms with natural interpretations. A unique feature of the proposed privacy mechanism is that it tells the truth while obfuscating the truth, which could help application domains that require information fidelity. Another desirable aspect of the privacy is its natural compatibility with unbounded data ranges, which has been a known challenge for local differential privacy.

- We develop fundamental properties of the proposed privacy mechanism, including the composition property that characterizes the level of overall privacy degradation under multiple queries of the same data, robustness to pre-processing and post-processing, and identifiability of the underlying data distributions. We also draw theoretical connections between interval privacy and the classical literature of survival analysis. The above theoretical developments also render the interval privacy a potential alternative or complement to existing privacy frameworks.

- We exemplify the use of interval privacy in estimating population distribution, statistical functional, and regression function, and show that the data collector does not necessarily need to know the distributional form of raw data for asymptotically accurate results. In particular, we develop a general method to perform supervised
regression with interval-privatized responses. The method can be applied to various interval-private data types, including pure intervals or a mixture of intervals and points. We provide theoretical justifications on how it works and practical guides to enhance computational efficiency.

- We demonstrate the proposed concepts, data formats, properties, and algorithms by various experiments. We also discuss the connections between interval privacy and local differential privacy from various angles. In particular, we show that 1) the above two notions of privacy have a trivial intersection, indicating a novel data privacy perspective brought by this work, 2) the interval mechanism can be regarded as an alternative type of limited-precision local privacy (Schein et al., 2019), and 3) the two notions of privacy may be generalized into a unified notion of local privacy.

1.1 Related work

We discuss some closely related literature in this subsection.

Data privacy. A popular framework of evaluating data privacy is through differential privacy (Dwork et al., 2006; Dwork, 2011), a cryptographically motivated definition of privacy that has gained significant attention over the past decade across different fields (Chaudhuri et al., 2011; Sarwate and Chaudhuri, 2013; Dong et al., 2019; Neunhoeffer et al., 2020; Vietri et al., 2020). Differential privacy measures privacy leakage by a parameter $\varepsilon$ that bounds the likelihood ratio of the output of a (private) algorithm under two databases differing in a single individual. A common tool for providing differential privacy is the sensitivity method (Dwork et al., 2006), which first computes the desired algorithm (e.g., sample median) on the data, and then adds noise proportional to the largest possible change induced by changing a single data entry in the database. Differential privacy is statistical database privacy which aims to protect the existence of individual identity in a database. The main difference between database privacy and local privacy (which we focus on here) is summarized below. First, local privacy aims to protect each data value, while database privacy protects whether an individual (data value) is included in the database or not. Second, local privacy is designed to disclose and collect individual-level data, while database privacy is developed for querying summary statistics. Third, database privacy often involves three parties: data creator (often individuals), a data collector (a trusted third-party, often large companies) who maintains the database, and analysts (often data scientists) who query statistics from the data collector; Local privacy may only involve two parties: a data creator and a data collector who may immediately analyze the collected data. Non-statistical data privacy frameworks include secure multi-party computing (Yao, 1982; Chaum et al., 1988) and homomorphic encryption (Gentry, 2009; Armknecht et al., 2015), which share privatized data using cryptographic methods. Cryptographic solutions are often lossless (in terms of encrypted information), but may need to be designed on a case-by-case learning scenario and often require a trusted third-party.

Local privacy. Existing solutions to local data privacy include cryptographic methods (Diffie and Hellman, 1976; Gentry, 2009), which depends on private keys to encrypt and decrypt data, and statistical methods which randomize raw data to guarantee a certain privacy level. Interval privacy can be regarded as a statistical framework for local data privacy. Closely related to our work is the framework of local differential privacy (Evfimievski
a local version of differential privacy (Dwork et al., 2006) where only randomized data are available to data collectors and analysts. Local differential privacy is a criterion that restricts the conditional distributions of the privatized data on any two different raw data to be ‘close’ (elaborated in Section 2.6). An information-theoretic counterpart of the concept was studied in (du Pin Calmon and Fawaz, 2012; Sun and Tay, 2019a). We exemplify the difference between interval privacy and local differential privacy through a toy example. Consider a salary of $25k and another salary of $250k. In a typical differential privacy scenario, the two private values are obfuscated with two random point values whose distributions exhibit epsilon-difference (often by adding noise). In contrast, under an interval privacy mechanism, the two private values are obfuscated with two random intervals, say $(0, $100k)$ and [$200k, \infty)$. The critical difference is that interval privacy offers information by narrowing down the support size and telling the truth, while local differential privacy offers information by perturbing the truth.

**Survival analysis.** From a theoretical perspective, the interval privacy is closely connected with classical survival analysis, where the survival time is known to fall into a certain inspection time interval (Kaplan and Meier, 1958; Turnbull, 1976). For example, patients’ infection time is often right-censored, meaning that it is either exactly observed or not observed but known to be larger than an inspection time. Though the interval data in survival analysis is naturally observed instead of artificially constructed, the existing theoretical machinery may be readily applied to estimate population-level information. The machinery includes Kaplan-Meier estimator of distribution functions (Kaplan and Meier, 1958) from right-censored data, maximum likelihood estimation of distribution functions from interval-censored data (Groeneboom and Wellner, 1992; Gentleman and Geyer, 1994; Geskus and Groeneboom, 1999), and simple linear regression with right-censored response (Miller, 1976; Buckley and James, 1979; Koul et al., 1981) and covariates (Gómez et al., 2003), among many others. Its theoretical root in survival analysis indicates a great potential for interval privacy as a general and flexible privacy technique for various application domains. A contribution of this work is to introduce interval-censoring as a general mechanism for interval privacy and study various theoretical properties in the context of data privacy.

**Signal quantization.** A related technique to random censoring is quantization (or deterministic censoring) that has been used in different ways in signal processing, including data-reduction parameter estimation for low-complexity sensors (Msechu and Giannakis, 2011), efficient online estimation of large-scale linear systems (Berberidis et al., 2016), one-bit compressive sensing (Boufounos and Baraniuk, 2008; Fu and Chi, 2018), efficient optimization (Alistarh et al., 2017), etc. The general purpose of quantization is to transform data $y$ into an element of a finite alphabet (which is usually a partition of the data space), but not to affect the estimability of a parameter unrelated to the distribution of $y$. For example, the problem of one-bit compressive sensing is to estimate linear models, $\beta$ in $y = \beta^T x + \varepsilon$, using $x$ and the sign of the response variable, sign($y$). This data may be regarded as a degenerate random censoring [$u, 1_{y>u}$] when $u$ is always zero. The distribution of $y$ (e.g., Gaussian) itself is not estimable given quantized observations because of the loss of distributional information in quantized regions, unlike a random censoring scenario where the distribution can be consistently estimated (Gentleman and Geyer, 1994). Nevertheless, the regression parameter in one-bit compressive sensing is still estimable, because of the sufficient infor-
information provided by distributional knowledge of the noise $\varepsilon$ so that a parametric likelihood function can be established (Fu and Chi, 2018). We will consider random censoring for data privacy so that individual data is probabilistically masked, but the population-wide distributional information is preserved (in case data analysts need it). We will revisit the regression problem with a random-censored response in Section 3.

2. Interval Privacy

This section introduces the notion of interval privacy, general interval mechanisms and data formats, and some fundamental properties of the proposed mechanisms.

2.1 Notation

Throughout the paper, we let $Y$ denote the random variable that represents the raw data. Suppose that the raw data $Y_1,\ldots,Y_n \in \mathcal{Y} \subset \mathbb{R}$ are continuously valued and i.i.d. with probability $p_Y$, density $p_Y$, and cumulative distribution function (CDF) $F_Y$.

In a local privacy scenario, data owners do not trust the data collector. A general local privacy scheme uses a random mechanism $M : \mathcal{Y} \rightarrow \mathcal{Z}$ that maps each $Y_i$ to another variable $Z_i \in \mathcal{Z}$, and then collects $Z_i$. The mechanism is often modeled by a conditional distribution $P_{Z|Y}(\cdot | Y)$. Let $p_Z$ denote the marginal density of $Z$. The random variable $Z$ may be constructed by a measurable function of $Y$, or a function of $Y$ and other auxiliary random variables. We assume that the joint distribution $[Y,Z]$ exists. Let $p_Z$ denote the marginal density of $Z$. The random variable $Z$ may be constructed by a measurable function of $Y$, or a function of $Y$ and other auxiliary random variables. We assume that the joint distribution $[Y,Z]$ exists. Suppose that $S$ is an interval or a union of intervals from where $Y$ takes value. We let $L(S) = p_Y(S)$ denote the “size” of the set $S$. The notion of $S$ could be generalized to an element of the Borel $\sigma$-algebra over $Y$.

2.2 Interval Privacy

**Definition 1 (Interval Privacy)** A mechanism $M$ has the property of interval privacy, if almost surely for all $y_1,y_2 \in S_z$,

$$
\frac{p_{Y|Z}(y_1 | Z = z)}{p_{Y|Z}(y_2 | Z = z)} = \frac{p_Y(y_1)}{p_Y(y_2)},
$$

(1)

where $p_{Y|Z}$ denotes the distribution of $Y$ conditional on $Z$ and $S_z$ is the support of $Y$ given $Z = z$.

The privacy coverage of $M$, denoted by $\tau(M)$, is defined by $\mathbb{E}(L(S_Z))$ (where the expectation is over $Z$). An $M$ is said to have $\tau$-interval privacy if $\tau(M) \geq \tau$.

**Implication 1**: We will give explicit formulas for some particular designs in Section 2.5. Condition (1) means that the conditioning on $Z = z$ does not provide extra information except that $y$ falls into $S_z$. If $y_1 \neq y_2$, and they fall into the same support $S_z$, then their likelihood ratio remains the same as if no action is taken. Condition (1) also implies that

$$
p_{Y|Z}(y | Z = z) = c_z 1_{y \in S_z} \cdot p_Y(y)
$$

(2)
Interval Privacy holds for the normalizing constant $c_z = 1/\int_{S_z} p_Y(y)\,dy$.

Implication 2: Suppose that $Y = y_1$ is to be protected, interval privacy creates ambiguity by obfuscating the observer with sufficiently many $y_2$’s in a neighborhood whose posterior ratios do not vary by incorporating the new information $Z = z$. Also, suppose that $S_Z$ is a (closed or open) interval, then the finite cover theorem implies the following alternative to the above second condition. For all $y$ in the interior of $S_z$, there exists an open neighborhood of $y$, $U(y) \subset S_z$, where (1) holds for all $y_1, y_2 \in U(y)$.

Implication 3: By its definition, the privacy coverage $\tau(M)$ takes values from $[0, 1]$. The privacy coverage quantifies the average amount of ambiguity or the level of privacy. A larger value indicates increased privacy. Likewise, for each pair of $[y, z]$ (nonrandom), we introduce $L(S_z)$ as the individual privacy coverage, interpreted as the privacy level for a particular data being collected (illustrated by the first column in Figure 1). To realize interval privacy, we will introduce a natural interval mechanism that converts $y$ to a random interval that contains $y$. In such a scenario, $S_z$ will be in the form of $[\ell, r]$, and thus $L(S_z) = r - \ell$.

Implication 4: A related measure is $1 - \tau(M)$ which naturally describes the privacy leakage. In an extreme case where the conditional support $S_z$ is the original support $Y$, we have $1 - \tau(M) = 0$, which means no privacy is leaked. We will show that within a large class of privacy mechanisms, the amount of leakage grows linearly with the number of observers (Theorem 8), a finer interval observation tends to reduce the privacy coverage (Theorem 12).

We provide Fig. 2 to visualize our unique approach to protecting data information. It shows the data format of interval data and its released information of the raw data as implied by posterior uncertainty. The approach is compared with the popular approach to privatizing data by perturbations. From a Bayesian perspective, the perturbation approach typically changes the density shape by more concentrating at a particular area; the interval approach will not change the density shape but restrict the essential support. Specifically, in the experiment, raw data $y$ was generated from the standard Gaussian. The interval privacy used standard Logistic random variable $U$ and reports either ‘$\leq U$’ or ‘$> U$’, with about 0.25 privacy leakage. The differential privacy budget truncated the raw data within $[-3, 3]$ and added Laplacian noises so that the privacy budget was $\alpha = 2$.

2.3 General Interval Mechanism and Data Format

A natural privacy mechanism to realize interval privacy is to generate a set of random ‘anchor’ points $Q = [Q_0, Q_1, \ldots, Q_m]$ with $-\infty = Q_0 = Q_1 < \cdots < Q_m = \infty$, and report the interval $(Q_{i-1}, Q_i]$ into which $Y$ falls.

In the general case, a raw data point $y \in \mathbb{R}$ is ‘observed’ as an interval that covers it, and a special case is when the interval degenerates to the exact point $y$. For multi-dimensional data, we will privatize each entry of it. Since a data analyst is usually interested in learning at the population level, we assume that the set of raw data points $y_1, \ldots, y_n$ are i.i.d. following a distribution function $F_y$ that is unknown. Let $Y$ denote a generic raw data to be privatized, so that $Y \sim F_y$. Under an interval privacy mechanism, the collected data is not $Y$ itself, but a privatized data $Z$ motivated by the following scenarios.

- A data owner will report an interval that contains each raw data.
A data owner will sometimes report its raw data once it falls into an ‘acceptable’ range.

Specifically, we generate \(Z\) in the following way. We partition the data domain \(\mathcal{Y} \subseteq \mathbb{R}\) into disjoint intervals, and report the interval which \(Y\) falls into. Occasionally, the point value \(Y\) is reported if it falls into a pre-specified range. The resulting data would be in the form of intervals, or a mixture of intervals and points. Formally, we introduce the following notion.

Let \(\{Q^{(1)}, \ldots, Q^{(m)}\}\) denote a partition of \(\mathcal{Y}\), where each \(Q^{(i)}\) is assumed to be a Borel set. For notational convenience, in the paper we suppose that \(Q^{(i)}\) is in the form of \((Q_{i-1}, Q_i]\) for \(i = 1, \ldots, m\), where \(Q_1 < \cdots < Q_{m-1}\) are randomly generated anchor points and \(Q_0 = \text{ess inf}\mathcal{Y}\) and \(Q_m = \text{ess sup}\mathcal{Y}\). Suppose that if \(Y\) falls into a pre-determined set \(A \subseteq \mathbb{R}\), named as the ‘acceptable set’, then the data owner chooses to disclose the value of \(Y\). In practice, the acceptable set is at the data owner’s discretion, and the set may not be fixed. To model the real-world complexity, we suppose that \(A\) can be one of the following: \(\emptyset\), a fixed set, or the union of \((Q_{k-1}, Q_k]\) for a fixed set of \(k\).
the form of $A$ is pre-specified and independent with $Y$. From Subsection 2.5 and afterward, we will elaborate on the $A = \emptyset$ case and show that population information is identifiable even without exact values of $Y$. More discussions on $A$ will be made in Remark 6.

We introduce the following privacy mechanism.

**Definition 2 (Interval Mechanism)** A privacy mechanism, denoted by $M : Y \mapsto Z$, maps $Y$ to the following privatized data $Z$,

$$Z = [Q, I(Q,Y), Y \cdot 1_{Y \in A}],$$

(3)

where $Q = [Q_1, \ldots, Q_{m-1}] \in \mathbb{R}^{m-1}$ is a random vector independent with $Y$, and $I : (Q,Y) \mapsto i$ is the indicator function defined by $Y$ falling into $(Q_{i-1}, Q_i]$ ($i = 1, \ldots, m$). The corresponding privacy coverage and privacy leakage follow Definition 1.

The validity of an interval mechanism in terms of Definition 1 is given by the following result.

**Theorem 3** A mechanism $M$ in Definition 2 satisfies the interval privacy.

**Remark 4 (Practical implementation)** In practice, a data-private information collection system involves two parties, a data owner (“Alice”) and a data collector (“Bob”). A general collection procedure is outlined as follows. First, a system designer, who may or may not be one of the two parties, define a way of generating $Q$. Such a generating process may be open-source implemented so that it is transparent to both parties. Second, the two parties agree on the use of the mechanism for data collection. Third, for each data point $Y$ of Alice, an instance of $Q$ is generated, and Alice needs to report the interval to Bob. Additionally, Alice may optionally report the exact value, but this is at Alice’s discretion. At the end, the set of data Bob collects consists of intervals and possibly some exact values (degenerate intervals).

**Remark 5 (Interpretation of the data)** The observables include a partition of $Y$ (by $Q$), the interval that $Y$ falls (by $I(Q,Y)$), and the value of $Y$ only if $Y \in A$ (by $Y \cdot 1_{Y \in A}$). The information obtained from the privatized data $Z$ is either an interval containing $y$ or the exact value of $y$. It can be seen that an interval mechanism does not distort the value of $y$. This property does not hold for popular approaches where perturbations are added to the raw data.

An alternative notion to $I(Q,Y)$ is to use $m$ indicator variables $1_{Y < Q_1}, \ldots, 1_{Y < Q_m}$ to represent where $Y$ is located at. By definition, $1_{Y \leq Q_i} = 1$ if $i \leq I(Q,Y)$ and $1_{Y < Q_i} = 0$ otherwise. The values of $Q$ are usually random so that it is possible to identify the population distribution of $Y$. This point will be elaborated in Subsection 2.5. As such, $Z$ conditional on $Y$ is also a random variable where the randomness comes from $Q$.

The choice of anchor points $Q$ will affect the privacy-utility tradeoff. Consider an extreme case where $m$ is sufficiently large. Then the collected interval becomes narrow and the privacy coverage tends to zero. In another case where $m = 1$ and $Q \in \mathbb{R}$ has a large variance, the interval is nearly $(-\infty, \infty)$ so that little utility is offered. The choice of $Q$ that attains the optimal utility-privacy tradeoff depends on the learning problem, and is left as future work.
Remark 6 (Interpretation of A)  Recall that whether Alice's reports a particular raw data value is at her discretion. The acceptable set A is a mathematical abstraction of such a possibility. In the Definition 2, an empty set A corresponds to the particular case where all observables are intervals. A random A means Alice's acceptable range varies among data due to, e.g., randomness by nature or some side information concerned by Alice. On the other hand, a deterministic nonempty set A means that Alice has a fixed acceptable range uniformly for all data points.

An alternative definition of the privacy leakage is \( L(A) \), meaning the probability of observing the exact value of \( X \). Compared with Definition 1, the leakage here does not consider the distribution of intervals not covered in \( C \). In the particular case \( A = \emptyset \), the privacy leakage here is zero, which is appealing as a quantization also provides information. If the raw data is multidimensional, the interval mechanism will be applied to each dimension, and the privacy coverage will be evaluated accordingly. A similar technique to privatize multi-dimensional variables is left as future work. On the other hand, if \( C = 0 \), then the value \( L(S) \) could be zero (when \( Y \in A \)) in evaluating the privacy coverage \( \tau(\mathcal{M}) = \mathbb{E}(L(S)) \). In general, the acceptable set \( C \) may or may not be random, depending on application domains or specific mechanism designs. An example will be provided in the next subsection.

We will elaborate on some specific uses of the interval mechanism in Section 2.5. The interval mechanism satisfies some desirable properties, which we will briefly mention in the next section.

2.4 Fundamental Properties of the Interval Mechanism

We are particularly interested in this mechanism not only because of its simplicity but also the following appealing composition property. Suppose that there are \( k \) interval private algorithms (or 'observers') each querying the same data with a mechanism \( \mathcal{M}_j : Y \mapsto Z_j = \{Q_j, I(Q_j, Y), Y \cdot 1_{Y \in A_j}, j = 1, \ldots, k \}. The ensemble mechanism, denoted by \( \bigoplus_{j=1}^k \mathcal{M}_j \), is an interval mechanism induced by intersections of \( Q_j \)'s and the union of \( A_j \)'s. In other words, \( k \) observers work together to narrow down the interval where a particular \( Y \) falls.

Composition. Suppose that there are \( k \) algorithms or observers, each using a mechanism to query information. These \( k \) observers may choose to collaborate to narrow down the range where a particular \( X \) belongs. This motivates the following ensemble mechanism, denoted by \( \bigoplus_{j=1}^k \mathcal{M}_j \), which is itself an interval mechanism.

Definition 7 (Ensemble Mechanism) The ensemble of two privacy mechanisms

\[ \mathcal{M}_i : X \mapsto Z = \{Q_i, I(Q_i, Y), X \cdot 1_{X \in A_i}\} \]

with \( i = 1, 2 \) is defined by

\[ \mathcal{M}_1 \oplus \mathcal{M}_2 : X \mapsto Z = \{Q_1 \oplus Q_2, I(Q_1 \oplus Q_2, Y), X \cdot 1_{X \in A_1 \cup A_2}\}. \]

where \( Q_1 \oplus Q_2 \) denotes the vector of all the anchor points from \( Q_1 \) and \( Q_2 \), and \( A_1 \cup A_2 \) denotes the union of two sets \( A_1, A_2 \). In general, the ensemble of \( k \) privacy mechanisms, denoted by \( \bigoplus_{i=1}^k \mathcal{M}_i \), is recursively defined by \( \bigoplus_{i=1}^k \mathcal{M}_i = (\mathcal{M}_1 \oplus \cdots \oplus \mathcal{M}_{k-1}) \oplus \mathcal{M}_k \) (\( k \geq 2 \)).
Theorem 8 (Composition Property) Let $\mathcal{M}_1, \ldots, \mathcal{M}_k$ be $k$ (ensemble) interval mechanisms as in Definition 2. Then

$$1 - \kappa(\bigoplus_{j=1}^k \mathcal{M}_j) \leq \sum_{j=1}^k (1 - \kappa(\mathcal{M}_j)).$$ (4)

An interpretation of the above theorem is that the privacy leakage of any ensemble mechanism is no larger than the sum of each of them. It is worth noting that $Q_j$’s may or may not be independent with each other, so communications between observers are allowed for this composition property to hold. In other words, this composition property holds even if the $k$ mechanisms are adaptively chosen.

Preprocessing. Suppose that $g : \mathcal{Y} \mapsto g(\mathcal{Y})$ is a measurable function on $\mathcal{Y}$. Suppose that an interval mechanism is applied to $g(\mathcal{Y})$ instead of $\mathcal{Y}$ itself, with the observation $\mathcal{M} : g(\mathcal{Y}) \mapsto Z = \{Q, I(Q, g(\mathcal{Y})), Y \cdot \mathbb{1}_{g(\mathcal{Y}) \in A}\}$. This is one-to-one corresponding to the following “pullback” privacy mechanism

$$\mathcal{M}_g : Y \mapsto Z_g = \{g^{-1}(Q), I(g^{-1}(Q), Y), Y \cdot \mathbb{1}_{Y \in g^{-1}(A)}\}. $$ (5)

The following result shows that if a $\tau$-interval private observation is made on a transformation of $\mathcal{Y}$, $g(\mathcal{Y})$, then the privacy coverage of the original data $\mathcal{Y}$ is not smaller than $\tau$, or the privacy leakage at the original data domain is no larger than $1 - \tau$. Furthermore, the $\mathcal{M}_g$ here may be regarded as $\mathcal{M}_j$ in Theorem 8, so it also holds for $k$ observers that may target transformations of $\mathcal{Y}$ instead of $\mathcal{Y}$ itself.

Theorem 9 (Robustness to Preprocessing) For any interval mechanism $\mathcal{M}$, it holds that $\tau(\mathcal{M}_g) \geq \tau(\mathcal{M})$, where the equality holds if and only if $g^{-1}(\tilde{y})$ has zero $F_y$-measure for all $\tilde{y} \in g(\mathcal{Y})$.

Postprocessing. The next result says that the privacy leakage of $\mathcal{Y}$ will not be increased by any subsequent processing using $Z$.

Theorem 10 (Robustness to Post-processing) Suppose that $\mathcal{M} : \mathcal{Y} \mapsto Z$ is an interval mechanism with $\tau$-interval privacy. Let $f : Z \mapsto W$ be an arbitrary deterministic or random mapping that defines a conditional distribution $W \mid Z$. Then $f \circ \mathcal{M} : \mathcal{Y} \mapsto [Z, W]$ also meets $\tau$-interval privacy.

The amalgamation of composition property and robustness permit modular designs and analyses of interval privacy mechanisms. If each system element is separately interval-private, so will be their combination or subsequent post-processing.

2.5 Examples of the Interval Data

We presented a general mechanism in Definition 2 to ensure interval privacy. In this section, we will exemplify it through two representative yet simple interval mechanisms. We call them Case-I and Case-II intervals, where the names follow the convention in survival analysis (Zhang and Sun, 2010). We will focus on one dimensional data $\mathcal{Y} \in \mathcal{Y} \subseteq \mathbb{R}$. Let $(U, V) \in \mathbb{R}^2$ be a random variable that satisfies $\mathbb{P}(U \leq V) = 1$. Suppose that $Y \in R$ is the
data that could have been observed, but privatized so that we observe an interval that it falls. We consider the following two mechanisms.

**Case-I interval mechanism:** Either \( Y \leq U \) or \( Y > U \) is observed. The observations are \( n \) i.i.d. copies of \( Z = [U, \Delta] \), where

\[
\Delta = 1_{Y \leq U}
\]

is an indicator variable.

**Case-II interval mechanism:** Either \( Y \leq U \), \( U < Y \leq V \), or \( Y > V \) is observed. The observations are \( n \) i.i.d. copies of \( Z = [U, V, \Delta, \Gamma] \), where

\[
\Delta = 1_{Y \leq U}, \quad \Gamma = 1_{U < Y \leq V},
\]

are indicator variables.

Suppose that the age of Alice is to be privatized. The above Case-I is perhaps the simplest interval mechanism where only the indicator of whether Alice being elder than a randomly generated age is reported. The Case-II is the simplest case that allows at least one bounded interval (e.g., within 20 to 40 years old). The following summarizes some characteristics of the privacy mechanism.

**Fidelity:** An interesting aspect of the proposed privacy mechanism is that it protects data by ‘telling no lies’ about the actual values. It creates a significant ambiguity with masks while maintaining essential side information to reconstruct population laws.

**Non-informativeness:** Conditional on the revealed support set, no extra information is revealed since the relative densities do not change. In other words, the only information provided by the privatized data \( Z \) about the protected data \( Y \) is the (often wide) interval that contains \( Y \).

**Identifiability:** Suppose that the support of \( U \) contains that of \( Y \). It has been shown under reasonable conditions that the distribution of \( Y \) can be consistently estimated from interval observations even if the underlying distribution is nonparametric (Groeneboom and Wellner, 1992). Fig. 3 (left plot) shows a parametric and nonparametric estimation of the distribution function \( F_Y \) where \( Y \) follows a standard Logistic distribution. The parametric estimation uses the standard maximum likelihood approach. The nonparametric estimation is based on the self-consistency algorithm (Turnbull, 1976) implemented in the ‘Icens’ R package (Gentleman and Vandal, 2010). Note that a random selection of \( U, V \) is essential. If they are deterministic values, the mechanism will not be able to accurately estimate the distribution of \( Y \), since one can always find a distinct distribution whose mass on the predetermined intervals coincide. Below is an example that covers a wide range of inference tasks.

**Example 1 (Functional Estimation)** Suppose a data analyst is interested in estimating a smooth functional \( K(F_Y) \) of the underlying distribution function \( F_Y \). A nonparametric estimator is \( K(\hat{F}_Y) \) where \( \hat{F}_Y \) is the nonparametric maximum likelihood estimator of \( F_Y \) with provable guarantees (Groeneboom and Wellner, 1992). Specifically, all moment functionals \( K : F_Y \rightarrow \int_y y^k dF_Y(y) \), or more generally, linear functionals in the form of \( K : F_Y \rightarrow \int_y \phi(y) dF_Y(y) \) can be estimated in this way.
Example 2 (Mean Estimation) Sometimes, a statistical functional may be directly estimated without the need of estimating $F_Y$. For example, suppose that the raw data are i.i.d. $Y_i \in [a, b]$ for $i = 1, \ldots, n$, with unknown mean $\mu$. The observations are $Z_i = [U_i, \Delta_i]$, $i = 1, \ldots, n$, from the Case-I interval mechanism with $U_i \sim \text{i.i.d. Uniform}[a, b]$. We provide the following estimator,

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \left( \Delta_i (2U_i - b) + (1 - \Delta_i)(2U_i - a) \right).$$

(6)

Proposition 11 shows that $\hat{\mu}_n$ is a consistent and unbiased estimator of $\mu$.

Proposition 11 The estimator in Example 2 satisfies $E(\hat{\mu}_n) = \mu$ and $\text{var}(\hat{\mu}_n) = O(n^{-1})$.

Achievability: The ambiguity as quantified by privacy coverage $\tau$ (in Definition 1) can be controlled by the distribution of $[U, V]$, or the number of intervals, e.g., two in Case-I and three in Case-II. The larger $\tau$, the more ambiguity and thus more protection. The following result shows that any level of privacy coverage between $[c, 1]$ for some constant $c \in [0, 1)$ is achievable by a suitable choice of $U, V$.

Theorem 12 Assume that the density function of $Y$ is bounded. For any $\tau \in (1/2, 1)$ (resp. $(1/3, 1)$) there exists a Case-I (resp. Case-II) mechanism $\mathcal{M}$ whose privacy coverage is exactly $\tau$.

The result and its proof indicate that a privacy mechanism exists for a privacy coverage arbitrarily close to one. Also, the proof indicates that the choice is not unique. As a byproduct of the proof, $\tau(\mathcal{M}) = n^{-1} \sum_{i=1}^{n} [F_Y(u_i) + (1 - F_Y(u_i))^2]$ is a consistent estimator of the privacy coverage for Case-I. And there is a similar estimator for Case-II. The above result indicates that we can achieve the max privacy near one, so why would not one always do so in practice? The practical concern is an inherent tradeoff between privacy coverage and statistical accuracy. To see this, we provide the following result, which is interesting in its own right.

For any functional that is differentiable along Hellinger differentiable paths of distributions (e.g. linear functionals), one can derive a Hájek-LeCam convolution theorem type information lower bound, giving the best possible limit variance that can be attained under $\sqrt{n}$ convergence rate where $n$ denotes the data size (Geskus and Groeneboom, 1999). The distribution of anchor points is said to be optimal if such information lower bound is attained by the produced interval data.

Theorem 13 An optimal distribution of $U$ (in the Case-I mechanism) for estimating any linear functional in Example 1 exists, and it has the density

$$g_U(u) = c_{\phi} \left\{ F_Y(u)(1 - F_Y(u)) \right\}^{1/2} \left| \frac{d}{du} \phi(u) \right|$$

if it is integrable, where $c_{\phi}$ is a normalizing constant.

The above result indicates a tradeoff between privacy coverage and statistical efficiency (in inference). Fig. 3 shows an example that illustrates the estimation of $F_Y$ and some optimal Case-I interval mechanism.
Figure 3: An illustration of the true CDF of the standard Logistic distribution, and its estimation using nonparametric and parametric methods with 1000 data (left plot), and the density of the standard Logistic distribution ($Y$), its optimal densities of the Case-I anchor ($U$) for estimating the first & second moments (right plot).

2.6 Discussions on the Local Differential Privacy and Interval Privacy

2.6.1 Local differential privacy and its relationship with interval privacy

A popular notation of privacy is the following local differential privacy (see, e.g., Evfimievski et al., 2003; Dwork et al., 2006; Kasiviswanathan et al., 2011; Duchi et al., 2013; Sarwate and Sankar, 2014).

**Definition 14 (Local Differential Privacy)** For a given privacy parameter $\alpha \geq 0$, a privacy mechanism $M$ is $\alpha$-differentially locally private if for all $y_1, y_2 \in \mathcal{Y}$,

$$\sup_{A \in \sigma(\mathcal{Z})} \frac{\mathbb{P}_{Z|Y}(z \in A \mid Y = y_1)}{\mathbb{P}_{Z|Y}(z \in A \mid Y = y_2)} \leq e^\alpha \tag{7}$$

where $\sigma(\mathcal{Z})$ denotes an appropriate $\sigma$-field over $\mathcal{Z}$.

Both the above privacy and interval privacy are local, suitable for scenarios where data collecting agents are untrustworthy. When the conditional densities exist, an equivalent condition of (7) is to require

$$\frac{p_{Z|Y}(z \mid Y = y_1)}{p_{Z|Y}(z \mid Y = y_2)} \leq e^\alpha \tag{8}$$

for all $z \in \mathcal{Z}$ and $y_1, y_2 \in \mathcal{Y}$ (almost surely). Suppose that a joint distribution of $Y, Z$ exists. By the Bayes’ theorem, (8) is further equivalent to

$$\frac{p_{Y|Z}(y_1 \mid Z = z)}{p_{Y|Z}(y_2 \mid Z = z)} \leq \frac{p_Y(y_1)}{p_Y(y_2)} e^\alpha. \tag{9}$$
The following result shows that interval privacy and local differential privacy do not imply each other, and their intersection is a trivial solution with null utility and maximal privacy (or $\tau = 1$ and $\alpha = 0$).

**Proposition 15** A privacy mechanism $M : Y \rightarrow Z$ that simultaneously satisfies $\tau$-interval privacy and $\alpha$-local differential privacy ($\alpha < \infty$) is trivial, meaning that $Z$ and $Y$ are independent.

An interesting problem is to quantitatively relate interval privacy and local differential privacy. Though interval privacy is neither weaker nor stronger than $\alpha$-differential privacy since, a possible way of relating these two is through privacy-utility tradeoffs. Specifically, we first record the privacy-accuracy tradeoff curve under each privacy framework and then map the two parameters ($\tau$ and $\alpha$) through the same learning performance on the curve. The above will provide a way to define ‘analogous parameters’ for interpretation and perception mathematically. In a numerical example, we generate 100 i.i.d. samples of $Y \sim \text{Uniform}[0, 1]$ and suppose that the distribution of $Y$ is unknown except that it falls into $[0, 1]$. We applied Case-I interval mechanism with $U \sim \text{Uniform}[-b, 1 + b]$ with $b = [22, 8, 5, 4, 2, 1.5, 1]$. We applied the technique in Example 2 to estimate $\mu = \mathbb{E}(Y)$. We measure the utility as $\mathbb{E}|\hat{\mu} - \mu|$, where $\mathbb{E}$ is approximated from 1000 independent replications. For comparison, we also used the $\alpha$-local differential privacy mechanism by perturbing $Y$ with Laplacian noises. We choose $\alpha = 0.2, 0.5, 0.8, 1, 2, 2.5, 3$ so that the utility under each $\alpha$ is almost the same as that under the counterpart $b$ of interval privacy. We visualize the ‘analogous parameters’ in Figure 4. We note that there exists no universal relationship between $\tau$ and $\alpha$, as the tradeoff curves depend on the underlying learning task. An interesting future direction is to use human perception (of privacy) as an evaluation criterion additionally to mathematical quantities such as $\tau$ and $\alpha$. 

![Figure 4: Experiments in Section 4.4: 1) the prediction error versus privacy coverage (left plot), and 2) privacy coverage versus the spread of intervals, as measured by the standard deviation of $U$ (right plot). The shaded bands indicate $\pm$ standard errors, from 50 independently replicated experiments.](image)
2.6.2 Interval privacy as a class of limited-precision differential privacy

To meet the local differential privacy, the condition in (8) needs to hold for any pair of original data \( y_1 \) and \( y_2 \). A privacy mechanism typically requires to add noise that scales with \( \max_{y_1, y_2 \in \mathcal{Y}} |y_1 - y_2| \) if \( \mathcal{Y} \) is moderately large, or to use other techniques such as asymptotic truncation (Duchi et al., 2018). The above concern motivated the notion of limited-precision local differential privacy (Schein et al., 2019), which generalizes the local differential privacy. A limited-precision counterpart of the differential privacy was also studied in (Flood et al., 2013).

**Definition 16 (Limited-Precision Local Privacy)** A privacy mechanism \( \mathcal{M} \) is \((d, \alpha)\)-locally private if for all \( y_1, y_2 \in \mathcal{Y} \) such that \(|y_1 - y_2| \leq d\), Condition (7) or (8) holds.

The above notion is suitable when the data owner only needs to hide the data within a given range of ambiguity, e.g., the count of certain words in an email instead of all the words (Schein et al., 2019), and the accurate location within a geographic range instead of all the world (Andrés et al., 2013). An equivalent way of expressing Definition 16 is that for any original data \( y_1 \) and any \( y_2 \in [y_1 - d, y_1 + d] \), their conditional density ratio is within \([e^{-\alpha}, e^{\alpha}]\). Given this, interval privacy may be regarded as a class of limited-precision differential privacy where the neighborhood range \([y_1 - d, y_1 + d]\) is replaced with a random interval containing \( y_1 \) and the budget becomes \( \alpha = 0 \). Note that such a limited range can be unbounded in our interval mechanisms. A subtle difference between limited-precision differential privacy with \( \alpha \neq 0 \) and interval privacy is that the former has the worst-case robustness against adversarial side information because its collected information is barely true. The above motivates a possible future direction to amalgamate interval privacy with local differential privacy, which we will briefly introduce in the remainder of this subsection.

2.6.3 Differential-Interval Privacy: a generalization of both worlds

Motivated by the form of interval privacy and local differential privacy (9), we introduce the following generalization.

**Definition 17 (Differential-Interval Privacy)** The same as Definition 1 except that the second condition (1) is replaced with

\[
\frac{p_{Y|Z}(y_1 \mid Z = z)}{p_{Y|Z}(y_2 \mid Z = z)} \leq \frac{p_Y(y_1)}{p_Y(y_2)} e^{\alpha}
\]

for \( y_1, y_2 \) in a range determined by \( z \). A mechanism \( \mathcal{M} \) is said to meet \((\tau, \alpha)\)-differential-interval privacy if \( \tau(\mathcal{M}) \geq \tau \).

Compared with Definition 1, the requirement in (10) is weaker as it involves a (typically small) parameter \( \alpha \). In the differential-interval privacy, any two conditional densities are only required to be equivocal for significant coverage of \( y \) instead of all the domain of \( y \). Thus, the \((\tau, \alpha)\)-differential-interval privacy generalizes both local differential privacy (which corresponds to \( \tau = 1 \)) and interval privacy (which corresponds to \( \alpha = 0 \)). A mechanism to realize the differential-interval privacy is to first perturb the original data \( Y \) to \( \tilde{Y} \) and then report a random interval that contains \( \tilde{Y} \). In principle, this relaxation supports a flexible design of private data collecting and learning procedures. Further study of the generalization is beyond the scope of the paper and is left as future research.
3. Regression with Private Responses

This section proposes a general approach to fit supervised regression using interval-private responses and develop theoretical justifications.

3.1 Formulation

Suppose that we are interested in estimating the regression function with $Y$ being the response variable and $X = [X_1, \ldots, X_p] \in \mathbb{R}^p$ being predictors. Suppose that $Y$ has been already privatized, and we only access an interval-private observation of $Y$, while the data $X$ is fully visible. This scenario occurs, for example, when Alice (who holds $Y$) sends her privatized data to Bob (who holds $X$) to seek Assisted Learning (Xian et al., 2020). The scenario also occurs when $Y$ has to be private while $X$ is already publicly available.

Following the typical setting of regression analysis, we postulate the data generating model $Y = f^*(X) + \epsilon$, where $f^*$ is the underlying regression function to be estimated, and $\epsilon \sim F_{\epsilon}$ is an additive random noise. We suppose that $X$ is a random variable independent of $\epsilon$, and that $F_{\epsilon}$ is a known distribution, say Gaussian or Logistic distributions. We will discuss unknown $F_{\epsilon}$ in Remark 20.

Suppose that Case-I or II intervals are used, then the data are $n$ observations in the form of $D_i = [u_i, \delta_i, x_i]^T$, or $D_i = [u_i, v_i, \gamma_i, x_i]^T$, $i = 1, \ldots, n$. The function $f^*$ is unknown. A parametric approach typically models $f^*$ by a linear function in the form of $Y = X^T \beta + \epsilon$, where $\beta \in \mathbb{R}^p$ is treated as an unknown parameter, and $\epsilon \sim F_{\epsilon}$. The above model includes parametric regression and nonparametric regression based on series expansion with bases such as polynomials, splines, or wavelets. Suppose a data analyst is to estimate $\beta$ from $D_n$. A classical approach is by maximizing the likelihood function which can be written as $\prod_{i=1}^{n} \left[ F_{\epsilon}(u_i - x_i^T \beta)^{\delta_i} \{ 1 - F_{\epsilon}(u_i - x_i^T \beta) \}^{1-\delta_i} \right]$.

Though the likelihood approach is principled for estimating a parametric regression, its implementation depends on the specific parametric form of the regression function $f$, and its extension to nonparametric function classes is challenging. In supervised learning, data analysts typically use a nonparametric approach, such as various types of tree ensembles and neural networks. However, the observables are intervals in the form of $[L, R]$, and existing regression techniques for point-valued responses ($Y$) cannot handle interval-valued labels. As such, we are motivated to “transform” the interval-data format into the classical point-data format, to enable direct uses of existing nonparametric regression methods and software packages.

3.2 Interval Regression by Iterative Transformations

Our main idea is to transform the data format from intervals to point values. Ideally, we hope to transform $[L, R]$ to a single point so that classical methods are readily applicable. We propose to use

$$\tilde{Y} = \mathbb{E}(Y \mid D, X).$$

as a surrogate to $Y$. This is motivated from the observation that $\tilde{Y}$ is an unbiased estimator of $f(x)$ for a given $X = x$, namely $\mathbb{E}(\tilde{Y} \mid x) = \mathbb{E}(Y \mid x) = f(x)$. It is worth noting that we
cannot use $\hat{Y} = \mathbb{E}(Y \mid D)$, because its expectation is $\mathbb{E}(Y)$ instead of $\mathbb{E}(Y \mid x)$, and thus this surrogates will produce a large estimation error.

In general, suppose that we choose a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ such that for all initial values $\hat{f}$ the sequence $(\hat{Y}_i)$ converges in probability. Based on the above arguments in (12), it is tempting to solve the following optimization problem

$$
\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(\hat{Y}_i, f(X))
$$

(13)

where $\hat{Y}_i = \mathbb{E}(Y \mid D, X) = f(X) + \mathbb{E}(\varepsilon \mid D, X)$.

(14)

The following result justifies the validity of using $\hat{Y}$ as a surrogate of $Y$ to estimate $f$, if the $\hat{Y}_i$ in (14) were computable. We define the norm of $f$ to be $\|f\|_{\mathcal{F}} = \mathbb{E}(f(X)^2)$. Suppose that the underlying regression function $f^* \in \mathcal{F}$ where $\mathcal{F}$ is a parametric or nonparametric function class with bounded $L_2(\mathbb{P}_X)$-norms. The next result shows that the optimal $f$ obtained by minimizing (13) with a squared loss $\ell$ is asymptotically close to the underlying truth $f^*$. The proof may be emulated to justify other loss functions, and it is left as future work. For notational convenience, we let $\mathbb{E}_n(\hat{Y} - f(X))^2 = n^{-1} \sum_{i=1}^{n} (\hat{Y}_i - f(X_i))^2$ denote the average squared loss.

**Theorem 18** Suppose that

$$
\sup_{f \in \mathcal{F}} \left| \mathbb{E}_n(\hat{Y} - f(X))^2 - \mathbb{E}(\hat{Y} - f(X))^2 \right| \to_p 0
$$

as $n \to \infty$. Then any sequence $\hat{f}_n$ that maximizes $\mathbb{E}_n(\hat{Y} - f(X))^2$ converges in probability to $f^*$, meaning that $\|\hat{f}_n - f^*\|_{\mathcal{F}} \to_p 0$. as $n \to \infty$.

In practice, however, the calculation of $\hat{Y}$ itself is unrealistic as it involves the knowledge of $f(X)$. In other words, the unknown function $f$ appears in both the optimization (13) and calculation of surrogates (14). The above difficulty motivates us to propose an iterative method where we iterate the steps in (13) and (14), using any commonly used supervised learning method to obtain $\hat{f}$ at each step.

The pseudocode is provided in Algorithm 1, where Case-I intervals are considered for brevity. Some details are further discussed in Remark 20. We provide a theoretical guarantee of its convergence for linear regression models. In practice, we set the initialization by $\hat{f}_0(x) = 0$ for all $x$. We experimentally found that Algorithm 1 is robust and works well for a variety of nonlinear models, including tree ensembles and neural networks. Examples are provided in Subsection 4.2.

**Theorem 19** Suppose that the underlying true regression function is $f^*(x) = x^T \beta^*$, $\beta^* \in \mathbb{R}^p$, and the least squares estimate $\hat{\beta}_0$ is used in (13). Assume that $\varepsilon$ is supported on $\mathbb{R}$ with $\mathbb{E}(\varepsilon) < \infty$, the CDF $F_\varepsilon(\cdot)$ of $\varepsilon$ is continuously differentiable. Also assume that $|U - f^*(X)|$ is almost surely bounded by a constant $B$, and $\sup_{v: |v| \leq B} \left| 1 + G_\varepsilon(v) p_\varepsilon(v)/(F_\varepsilon(v)(1 - F_\varepsilon(v))) \right| < 1$, where $p_\varepsilon$ is the density of $\varepsilon$ and $G_\varepsilon(v) = \mathbb{E}(\varepsilon | X = x)$. Then there exists a constant $\delta > 0$ such that for all initial values $\hat{\beta}_0$ satisfying $\|\hat{\beta}_0 - \beta^*\| < \delta$, we have $\lim_{n,k \to \infty} \|\hat{\beta}_k - \beta^*\| = 0$ in probability.
Algorithm 1 Regression from Interval Response by Iterative Transformations (an Case-I example)

**Input:** Interval-valued responses $D_i$ in Eq. (11) and predictors $x_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, function class $F$, error distribution function $F$.

**Initialization:** Round $k = 0$, function $\hat{f}_0(\cdot)$

1: repeat
2: Let $k \leftarrow k + 1$
3: Update the representative $\tilde{y}_i$, $i = 1, \ldots, n$
   $\tilde{y}_i = \hat{f}_{k-1}(x_i) + \mathbb{E}(\varepsilon | \tilde{u}_i, \delta_i, x_i)$ (15)
   where $\tilde{u}_i = u_i - \hat{f}_{k-1}(x_i)$.
4: Fits a supervised model $\hat{f}_k$ using $(\tilde{y}_i, x_i)$ as labeled data by optimizing (13) using a preferred method
5: until Stop criterion satisfied (e.g. if the fitted values no longer change much)

**Output:** The estimated function $\hat{f}_k : \mathbb{R}^p \rightarrow \mathbb{R}$

**Remark 20 (Computation of the conditional expectation in Equation (15))** The term $\mathbb{E}(\varepsilon | \tilde{u}_i, \delta_i, x_i)$ is essentially a conditional mean indicated by $\delta_i$, or $\mathbb{E}(\varepsilon | \varepsilon \leq u_i - \hat{f}_{k-1}(x_i))$ if $\delta_i = 1$ and $\mathbb{E}(\varepsilon | \varepsilon > u_i - \hat{f}_{k-1}(x_i))$ otherwise. In implementing Algorithm 1, we need to specify a distribution for the error term $\varepsilon$ when calculating (15). In general, the learning accuracy can be suboptimal if a data analyst postulates a model that violates the underlying data generating process, an issue known as model mis-specification (Ding et al., 2018). Nevertheless, in our context of data-private regression, we found from various experiments that the accuracy of estimating regression functions is not sensitive to mis-specification of the noise distribution. We conjecture that the consequent estimation of the regression function is still consistent under mild conditions. A practical suggestion to data analysts is to treat $\varepsilon$ as Logistic random variables to simplify the computation. Discussions on the fast computation are included in the Appendix, along with some related experimental studies in Subsection 4.5. Also, if the standard deviation of the noise term $\sigma$ is unknown in practice, we suggest estimate $\sigma^2$ with $n^{-1} \sum_{i=1}^n (y_i - \tilde{y}_i)^2$ at each iteration of Algo 1.

4. Experiments

In this section, we provide some representative experiments to demonstrate the use of interval privacy, including an unsupervised example, a supervised one, and a real-data one. We will use Logistic-distributed anchor points since they can be easily generated from $\log \{U/(1 - U)\}$, where $U$ is a uniform random variable.

4.1 Estimation of Moments

In this experiment, we demonstrate interval-private data (with reasonably broad privacy coverage) in moment estimation. Suppose that 100 raw data are generated from $Y \sim \mathcal{N}(0.5, 1)$, and the private data we collect is based on the Case-I mechanism with $U \in \text{Uniform}[-T, T]$, where $T = 2n^{1/3}$. The goal is to estimate $\mathbb{E}(Y)$. We demonstrate three methods and compare them with the (unrealistic) estimates based on the raw data in Table 1. The first method, denoted by ‘Example 2’, uses the estimator in (6). By a similar argument as the proof of Proposition 11, the choice of $T$ guarantees that the $\mu$ can be consis-
Table 1: Mean absolute errors of estimating $\mathbb{E}(Y)$ and $\mathbb{E}(Y^2)$ using interval-private data and raw data (in hindsight), where $Y \sim \mathcal{N}(0.5, 1)$, subject to 0%, 1%, and 5% outliers. Errors were calculated from 1000 independent replications so that the standard errors are all within 0.01.

<table>
<thead>
<tr>
<th></th>
<th>Estimate $\mathbb{E}(Y)$</th>
<th>Estimate $\mathbb{E}(Y^2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n$</td>
<td>0%</td>
</tr>
<tr>
<td>Private data</td>
<td>100</td>
<td>0.81</td>
</tr>
<tr>
<td>(Example 2)</td>
<td>1000</td>
<td>0.77</td>
</tr>
<tr>
<td>Private data</td>
<td>100</td>
<td>0.53</td>
</tr>
<tr>
<td>(NPMLE)</td>
<td>1000</td>
<td>0.43</td>
</tr>
<tr>
<td>Private data</td>
<td>100</td>
<td>0.52</td>
</tr>
<tr>
<td>(MLE)</td>
<td>1000</td>
<td>0.43</td>
</tr>
<tr>
<td>Raw data</td>
<td>100</td>
<td>0.75</td>
</tr>
<tr>
<td>(mean)</td>
<td>1000</td>
<td>0.75</td>
</tr>
<tr>
<td>Raw data</td>
<td>100</td>
<td>0.74</td>
</tr>
<tr>
<td>(median)</td>
<td>1000</td>
<td>0.74</td>
</tr>
</tbody>
</table>

The second method, denoted by ‘NPMLE’, uses the nonparametric maximum likelihood approach (Turnbull, 1976) available from the ‘Icens’ R package (Gentleman and Vandal, 2010). The third method, denoted by ‘MLE’, presumes that the distribution is Gaussian with unknown mean and variance, and estimates the mean with the standard maximum likelihood principle. Unlike the above two methods, the third one requires the knowledge of a parametric form and is thus expected to converge faster. We also consider two methods based on the raw data in hindsight: the sample average and the sample median. To demonstrate the robustness of interval data, we add none, 1%, and 5% proportion of outliers (where $Y$ is set to be 999). We also considered the estimation of $\mathbb{E}(Y^2)$, where we used a similar procedure to collect interval-private $Y^2$, except that $U \sim \text{Uniform}[0, 2T]$ so that the privacy coverage remains to be about 0.95 (or 5% leakage).

Table 1 summarizes the results, which indicate that the estimation under highly private data is reasonably well when compared with the oracle approach (under 0% outlier). The interval-private data tend to be more robust against outliers than the estimation based on simple mean and comparable to the median (based on raw data).

4.2 Estimation of Regression Functions

In this experiment, we demonstrate the method proposed in Section 3 on the linear regression model $Y = f(X) + \varepsilon$, where $f(X) = \beta X$ is to be estimated from the Case-I privatized data $Z = [U, V, \Delta]$. We generate $n = 200$ data with $\beta = 1$, $\varepsilon \sim \mathcal{N}(0, 1)$, $U$ a Logistic random variables with scale 5. The corresponding privacy coverage is around 0.9. Fig. 5 demonstrate a typical result. The prediction error is evaluated by mean squared errors $\mathbb{E}(f(\tilde{X}) - \hat{f}(\tilde{X}))^2$ where $\tilde{X}$ denotes the unobserved (future) data. With a limited size of data, the algorithm will produce an estimate $\hat{f}(x) = \hat{\beta}X$ that converges well within 20 iterations. The initialization is done by simply setting $\hat{f}_0(x) = 0$. 


Figure 5: Experiments in Section 4.2: Snapshots of the method proposed in Section ?? for linear regression at the 1st (left-top), 3rd (right-top), and 20th (left-bottom) iterations, and the prediction error ($L_2$ loss) versus iteration (right-bottom). Grey vertical segments indicate the observed intervals in the form of $(-\infty, u]$ or $(u, \infty)$; Blue dots and lines indicate the unprotected data $Y$ and the underlying true regression function; Red dots and dashed lines indicate the adjusted data $\tilde{Y}$ in (12) and the estimated regression function.

In another experiment, we demonstrate the method proposed in Section 3 on the nonparametric regression model $Y = f(X) + \varepsilon$, using the Case-II privatized data $Z = [U, V, \Delta, \Gamma]$. Suppose that $n = 200$ data are generated from quadratic regression $f(X) =$
$X^2 - 2X + 3$, and $\varepsilon \sim \mathcal{N}(0, 1)$. Let $U = \min(L_1, L_2)$, $V = \max(L_1, L_2)$, where $L_1, L_2$ are independent Logistic random variables with scale 5. The corresponding privacy coverage is around 0.9. Random Forest (depth $3$, 100 trees) with features $X_1 = X$, $X_2 = X^2$ are used to fit Algorithm 1. Fig. 6 demonstrates a typical result. With a limited size of data, the algorithm will produce a tree ensemble estimate that converges well within 20 iterations.

### 4.3 Data study: Life Expectancy regression

In the experimental study, we considered the ‘life expectancy’ data from the kaggle open-source dataset (Kaggle, 2020), originally collected from the World Health Organization (WHO). The data consists of 193 countries from 2000 to 2015, with 2938 data items/rows uniquely identified by the country-year pair. The learning goal is to predict life expectancy using 20 potential factors, such as demographic variables, immunization factors, and mortality rates.

We will exemplify the use of Algorithm 1 under three mechanisms. The first mechanism (‘Oracle’) uses the raw data of $Y$ (life expectancy). The second mechanism (‘$M_1$’) is described $Y \rightarrow Z$, where $Z$ is in the form of (3) in Definition 2, where $Q = [U - 1, U + 1]$ with $U$ generated from the Logistic distribution with standard deviation $s = 1$, and $A = \{y : U - 1 < y \leq U + 1\}$. The interpretation is that during the data collection/survey process, individual data items with an overly short or long life expectancy tend to report a half-interval (either $\leq U - 1$ or $> U - 1$), while those within the mid-range $A$ tend to report the exact values without privacy concerns. The second mechanism (‘$M_2$’) is a Case-II mechanism described by (3), where $Q = [U, V]$ is generated from the ordered Logistic distribution with standard deviation $s = 2$. The third mechanism (‘$M_3$’) is a Case-I mechanism described by (3), where $Q = U$ is generated from the Logistic distribution with $s = 5$. The interpretation of $M_2$ or $M_3$ is that individual data items are quantized into two or three random categories. We calculated the privacy coverage (Definition 1) for each privacy mechanism using the empirical distribution and summarize it in Table 2.

For each mechanism, the predictive performance of the fitted regression under methods, linear regression (LR), gradient boosting (GB), and random forest (RF), are evaluated using the five-fold cross-validation. The performance results are summarized in Table 2. The results show that a privacy mechanism with smaller privacy coverage tends to perform better, which is the expected phenomenon due to privacy-utility tradeoffs. The results also show a (statistically) negligible performance gap between $M_1$, $M_2$, and the Oracle. The performance starts to degenerate only in the last mechanism, where there is a large privacy coverage (94%) or small privacy leakage (6%).

To visualize the data and individual-level privacy coverage, we also show a snapshot of the database in Figure 1. We used the Case-II mechanism and generated $U, V$ from the standard Logistic distribution.

### 4.4 Tradeoff between Learning and Privacy Coverage

The fundamental tradeoff between privacy coverage and learning performance is often computable in parametric learning, where the asymptotic variance and coverage privacy are treated as functionals of the distribution of $U$ (or $[U, V]$), and in some nonparametric learning contexts (see e.g., Theorem 13 and relevant discussions). In this experiment, we...
Figure 6: Experiments in Section 4.2: Snapshots of the method proposed in Section ?? for nonlinear regression at the 1st (left-top), 3rd (right-top), and 20th (left-bottom) iterations, and the prediction error ($L_2$ loss) versus iteration (right-bottom). Grey vertical segments indicate the observed intervals in the form of $(-\infty, u]$, $(u, v]$, or $(u, \infty)$; Blue dots and lines indicate the unprotected data $Y$ and the underlying true regression function; Red dots and yellow dashed lines indicate the adjusted data $\tilde{Y}$ in (12) and the estimated regression function.

demonstrate the tradeoff using $n = 200$ data as was generated in Subsection 4.2. We adopted Case-I intervals with $U$ generated from Logistic distributions with scale parameter $0.1, 0.3, 0.5, 1, 3, 5, 10, 20, 30$, and numerically compute the prediction error and privacy cov-
Table 2: Predictive performance of linear regression (LR), gradient boosting (GB), and random forest (RF) methods under different interval mechanisms. Evaluation are based on the empirical $R^2$ and mean absolute error (MAE) from five-fold cross validations (with standard errors in the parentheses).

<table>
<thead>
<tr>
<th>Coverage</th>
<th>$\mathcal{M}_1$</th>
<th>$\mathcal{M}_2$</th>
<th>$\mathcal{M}_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle</td>
<td>0%</td>
<td>57%</td>
<td>76%</td>
</tr>
<tr>
<td>LR</td>
<td>$R^2$</td>
<td>0.79(0.02)</td>
<td>0.78(0.02)</td>
</tr>
<tr>
<td></td>
<td>MAE</td>
<td>3.21(0.07)</td>
<td>3.25(0.08)</td>
</tr>
<tr>
<td>GB</td>
<td>$R^2$</td>
<td>0.89(0.01)</td>
<td>0.86(0.01)</td>
</tr>
<tr>
<td></td>
<td>MAE</td>
<td>2.3(0.18)</td>
<td>2.56(0.11)</td>
</tr>
<tr>
<td>RF</td>
<td>$R^2$</td>
<td>0.82(0.02)</td>
<td>0.78(0.02)</td>
</tr>
<tr>
<td></td>
<td>MAE</td>
<td>2.86(0.12)</td>
<td>3.25(0.09)</td>
</tr>
</tbody>
</table>

Figure 7: Experiments in Section 4.4: 1) the prediction error versus privacy coverage (left plot), and 2) privacy coverage versus the spread of intervals, as measured by the standard deviation of $U$ (right plot). The shaded bands indicate ±standard errors, from 50 independently replicated experiments.

The results, summarized in Fig. 7, indicate that the prediction error is not sensitive to privacy coverage unless it is very close to one.

4.5 Sensitivity of Mis-specified Noise

From our various experiments, the learning accuracy is generally not much hindered by a wrongly specified distribution of $\varepsilon$ when calculating (15) (in the regression context). We demonstrate the sensitivity of wrongly specifying a distribution term using a specific example. A more sophisticated sensitivity analysis is left as future work.
Figure 8: Performance (mean squared error) versus mis-specification level (in terms of the $t$-degree of freedom $d$). Larger $d$ means that less mis-specification. The shaded bands indicate ±standard errors, from 200 independently replicated experiments.

Data are generated in the same way as in Subsection 4.2, except that the true noises follow $t$-distributions with degrees of freedom $d = [1000, 100, 10, 5, 3, 1]$. Here, $d = 1000$ is virtually Gaussian while $d = 1$ corresponds to a (heavy-tailed) Cauchy distribution. The postulation is still a Gaussian noise (so that it is misspecified). The results summarized in Fig. 8 indicates that the learning performance (as evaluated by mean squared error) is not severely affected, and less deviation tends to produce less degradation in performance.

5. Conclusion and Further Remarks

In this work, we proposed concepts, properties, and some usages of interval privacy. We showed a composition property of the privacy leakage and exemplified the interval mechanism through some specific designs. In the context of supervised learning, we also proposed a method to learn from interval-valued data, which enables direct use of existing techniques for point-valued data, with provable guarantees.

Future work includes the following directions. First, it is worth extending the notion of interval privacy for continuously-valued variables to discrete variables such as categorical, ordinal, or counts, and from one-dimensional intervals to multi-dimensional subsets. Second, it would be interesting to extend existing learning algorithms or develop new algorithms to tackle interval-private data on a case-by-case basis. Third, a more sophisticated analysis of the tradeoff between privacy and learning utility deserves further study. Fourth, the proposed regression method addresses only privatized $y$. In many application scenarios, sensitive information is also stored in $X$, e.g., salary, gender, race, etc. A future direction is to develop general methods for supervised learning with fully private data.
Acknowledgements

The authors thank Xuan Bi, Yuejie Chi, Zhiwei Steven Wu, and Robert Calderbank for insightful discussions that have improved this work.
References


X. Xian, X. Wang, J. Ding, and R. Ghanadan. Assisted learning: A framework for multi-
